3-2017

Rethinking institutional repositories

Chee Hsien Aaron (ZHENG Zhixian) TAY
Singapore Management University, aarontay@smu.edu.sg

Rethinking Institutional Repositories

By Aaron Tay

Are institutional repositories (IRs) a dead end? Given how librarians have been ripping into them recently—and not so recently—you might think so. Criticisms of IRs go back as far as 2008 when Dorothea Salo wrote a scathing article on their management, titled “Innkeeper at the Roach Motel” (*Library Trends*, v. 57 No. 2, 2008: pp. 98-123). In her view, institutional repositories are almost always doomed due to a lack of support and no clear compelling vision, which is further handicapped by horrible repository software.

More recently, Eric Van de Velde seems more than ready to bury them. His July 24, 2016 post, “Let IR RIP” to the *SciTechSociety* blog calls IRs “obsolete.” He urges his readers to phase them out and consider other alternatives to support the Green road to open access (scitechsociety.blogspot.sg/2016/07/let-it-rip.html).

As recently as November 2016, George Macgregor, Institutional Repository Coordinator, University of Strathclyde, wrote that such rumblings are not new and have been expressed in person at open access conferences. Yet at Strathclyde, IRs still experience a high volume of deposited full text and digital objects. He traces various criticisms of IRs and notes that changing academic behavior “takes a long time.” He concludes that “IRs need to remain a principal mechanism for achieving Open Access, whether we like it or not.” (“The long read: Why do institutional repositories remain one of the only viable options for Green Open Access?”; strathoa.tumblr.com/post/152680188170/the-long-read-why-do-institutional).

Richard Poynder, in his September 2016 commentary that preceded his interview with Coalition for Network Information CEO Cliff Lynch, argues that while open access seems inevitable, he sees a “growing sense that green OA has lost its way” (richardpoynder.co.uk/Clifford_Lynch.pdf). He describes most IRs as full of entries with no full text, only bibliometric details. This is despite a flood of mandates from institutions and funders that apparently do not motivate researchers to self-deposit with institutional repositories.

As a librarian on the ground who has a keen interest in the area and who interacts with faculty and students regularly, I must say such pessimism is not entirely unfounded.

RELUCTANT RESEARCHERS

Some of the obvious reasons why researchers are reluctant to deposit their work in an IR includes ignorance of the existence of the IR and of their rights to self-archive, along with a lack of motivation because the current academic structure does not provide incentives for making papers open access. As someone who has worked in institutions that have a mandate and ones that have not, I can affirm that in general while open access mandates may help (depending on the

type of mandate), they aren't a silver bullet in getting researchers to voluntarily submit their work into institutional repositories.

I think more troubling is the rise of a group of researchers who actually are ready and willing to self-archive but choose to do it elsewhere rather than in the IR.

Increasingly, we see researchers who are motivated to self-archive their papers in both subject repositories or preprint repositories such as arXiv (arxiv.org), SSRN (ssrn.com), or in so called Scholarly collaboration networks (SCNs)/social sharing networks such as ResearchGate (researchgate.net), Academia.edu, and perhaps Mendeley (mendeley.com). Yet they are stubbornly lukewarm towards IRs.

I believe it's instructive to consider why researchers are choosing such sites to archive their papers over IRs despite the obvious drawbacks of the other sites. Why not IRs? Here are some reasons why.

Institutional affiliations are not permanent leading to lack of ownership
LACK OF PERMANENT OWNERSHIP
Many, if not most, researchers tend to change institutions at least once in their careers. This impermanence of institutional affiliation leads a lack of ownership, particularly if you include their time as a PhD student. The main attraction of creating profiles or accounts at subject repositories like SSRN or in SCNs like ResearchGate is that researchers will always have control of that account and control of the papers they deposit in those venues even if they change institutions. It's no surprise that researchers tend to have a sense of ownership over such accounts.

While ORCID (orcid.org) is posed to eventually diminish the impact of such issues by allowing researchers to own one unique author identifier throughout their career while pushing information to various research profiles, the full text of the paper has to sit somewhere.

Researchers who put their papers in an IR will eventually lose direct control of those papers when they leave the institution and are unable to easily edit or make changes to those papers. By putting all their papers in one central source, they retain control.

They can also obtain aggregated usage statistics in one place, as compared to having usage statistics spread around in various IRs, which will be unwieldy to aggregate—assuming you could even aggregate statistics that are not standardized across various repositories.

SUBJECT STABILITY
The flip side of the fact that institutional repositories are often not permanent is that subject/discipline affiliations most likely are stable. Researchers may move

from one institution to another, but if they are researching in a particular discipline, such as history or physics, they will probably continue in that subject area.

Subject Repositories have the advantage of greater familiarity to scholars and can have systems custom built for each researcher's community. Subject repositories and/or preprint servers have the advantage because researchers tend to think along disciplinary lines and in many disciplines there is already a tradition of putting up preprints prior to publication.

By putting papers in subject repositories like SSRN, researchers can benchmark their paper against their peers in the same discipline, something that is not possible in IRs.

<insert screenshot of SSRN most downloaded papers/authors>

Given the central mass of disciplinary appropriate eyeballs already there, it's no surprise that IRs tend to lose out to subject repositories in terms of interest. As the saying goes, out of sight is out of mind.

INCONSISTENCIES AND POOR USER EXPERIENCES

Institutional repositories tend to have poor user experience and are inconsistent from one another. It's fairly well known that, compared to SCNs, most institutional repositories lag behind in functionality and sophistication. For example, until recently most institutional repositories did not automatically pull metadata to ease the task of entering bibliometric data, nor did they automatically do checks on SHERPA/RoMEO (sherpa.ac.uk/romeo) and send out emails to researchers to inform them that a paper they published could be self-archived.

In comparison, ResearchGate and Academia.edu are constantly innovating. Although many people find them very spammy and intrusive, I think they do at least try to use the latest known gamification and social networking techniques to try to encourage use.

For example, ResearchGate can tell you who viewed your record, who downloaded and read your paper (if they were signed on while doing so), and you can even respond to such information by asking the identified readers for a review.

<insert screenshot of ResearchGate shows who downloaded your paper>

Not everyone considers such features to be positive, but the point here is that they are iterating much quicker than the average institutional repository.

Even the lack of new innovative features may not be the only reason for the failure of IRs, rather it is the lack of consistency between institutional repositories.

Even though most University IRs are using a relatively small set of common software—Digital Commons (digitalcommons.bepress.com), DSpace (dspace.org), or EPrints (eprints.org)—they can vary greatly depending on the customization and feature set.

For the already time strapped researcher, learning to come to grips with a new system (with different submission formats, interface, and requirements) whenever they change institutions seems to be too much work, particularly when they have alternatives.

In fact, scholarly communication librarians have given up on getting researchers to submit papers on their own and gone with the mediated deposit model where they upload papers on their behalf. Many librarians also trawl the web looking for other papers archived by their researchers at other sites such as subject repositories or ResearchGate. But is that going to scale when repositories have poor interoperability and traditionally been built to support individual researcher uploads and not bulk uploading?

THE QUESTION OF MASS

In the last 10 years, we have learned that having mass on the web is important and network effects tend to dominate. This results in giants such as Facebook that are almost impossible to dislodge even with titanic efforts from companies such as Google. Facebook became too entrenched due to network effects.

Will ResearchGate and its peers that aim to be the academia equivalents of Facebook succeed using the same centralized, walled garden strategy? We know that many of the social and networking aspects that ResearchGate and to a lesser extent subject repositories like PubMed and SSRN bring are nearly impossible to replicate on isolated siloed IRs.

We know that IRs today are not destinations for visitors—most visitors discover papers on our repositories via discovery search engines like Google Scholar which link them directly to the PDF. Very few see the actual repository software pages except for the few brave souls that submit papers. This in itself isn't an issue if the aim is just open access, however it does prevent the social network effect from occurring, since it seems the first step of getting researchers to care about depositing papers in your site is to get them to come to your site in the first place!

AGGREGATING DATA

One way to counter the lack of mass of individual systems is to allow aggregation of each IR. While aggregator systems, such as CORE (core.ac.uk), BASE(base-search.net), and SHARE (share-research.org), exist to attempt to aggregate all data into a centralized repository, the lack of standardization among repositories in terms of consistency of metadata makes the whole aggregator process a little pointless, particularly with outdated protocols like OAI-PMH.

This is also where the earlier weaknesses of the lack of consistency and standards among repositories rears its head. It's not just surface usability and features that differ between one institutional repository and another, but also the limitation of almost no standards for metadata and content. It's becomes a mishmash of formats when you try to search across them using aggregator systems.

Even something as simple as identifying whether an entry harvested via OAI-PMH has full text attached is a nightmare. Incidentally, this is a reason why most libraries using web scale discovery systems often do not include IR contents from outside their university. Many of the contents in IRs are in fact indexed in web scale discovery systems, including Summon (proquest.com/products-services/The-Summon-Service), Primo (exlibrisgroup.com/category/PrimoOverview), and EBSCO Discover Service (ebscohost.com/discovery). however, most discovery system managers prefer not to turn them on because this results in many items surfacing that can't be reliably marked as full text, leading to a lot of confusion.

The exception to this rule is the Digital Commons network by bepress (digitalcommons.bepress.com), which shows what can be achieved by ensuring repositories have a constant set of standards. By using the cloud based digital commons repository, and assuming you keep the recommended subject scheme, you can easily compare the usage of items on your repository versus other repositories on the same network in the same discipline. It also has no issue detecting which items are full text and which aren't.

<insert screenshot of digital commons network showing most downloaded paper by discipline across institutions>

For an example of how the lack of consistency hurts, I was studying oaidoi (oadoi.org), a nifty new service that allows you to feed it a digital object identifier (DOI) and it would try to see if a postprint version exists on a repository by checking the BASE aggregator (among other sources).

While this works fine in theory, I've found it can fail for various reasons such as the repository not assigning DOIs to post prints or in other cases it simply does not expose the DOI to the BASE harvester. A consistent standard would help greatly here.

THE RISE OF REPOSITORIES 2.0?

IRs have a natural advantage over centralized silos in that they are less easily taken over or disrupted. The recent purchase (May 2016) of SSRN by Elsevier is a good example of the vulnerability of centralized repositories. But beyond that, there are of course defenders of repositories who rally the repository crowd by mooting the idea of next generation repositories that overcome the weaknesses I

mentioned earlier. Chief among them is Kathleen Shearer, Executive Director of the Confederation of Open Access Repositories (COAR). While she admits that repositories haven't been as successful as hoped, she believes the answer is to work on the flaws of repositories to improve on them and not to give up (coar-repositories.org/news-media/more-on-the-future-of-repositories-response-to-richard-poynder).

COAR (coar-repositories.org) has launched various initiatives and working groups to address many of the issues, including working on guidelines for repository interoperability, standardizing Controlled Vocabularies for Repository Assets, and studies on metrics such as usage. Coupled with work on new protocols to replace the aging OAI-PMH standard and discussions into value added services that repositories could serve beyond being just repositories of content, this is what I believe constitutes the next generation repositories.

The hope is that such next generation repositories will be interoperable, and serve as knowledge nodes in the scholarly system that can be seamlessly aggregated to counter the mass of centralized repositories. Already we see the rise of regional repository networks like LA Referencia (lareferencia.info/) and OpenAIRE (openaire.eu) leading the way.

Will these initiatives take off? Only time will tell. But I hope they do.

**********BOX**********

Reasons Why Researchers Avoid IRs

Institutional affiliations are not permanent, leading to lack of ownership.

Subject/Discipline affiliations are stable and subject repositories might be a more natural level of aggregation

Institutional repositories tend to have poor user experience and are inconsistent from one another